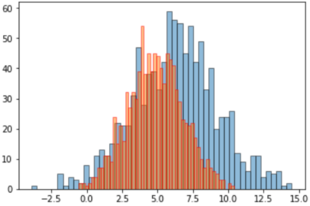
Next arguments

Cross entropy

Loss functions

What loss functions should we use for comparing probability distributions?



We could treat them as “normal functions”, and use e.g. quadratic distance between true and predicted probabilities.

Can we do better? For instance, in logistic regression we do not use mean squared error, but use negative loglikelihood. Why?

The likelihood of data

Let us consider a training set composed by data $\langle x_i, y_i \rangle$. y_i is here the true lable associated with the input x_i.

We have two distributions of interest: the true (categorical) distribution

$$P(y|x_i) = \begin{cases} 1 & if\, y = y_i \\ 0 & otherwise \end{cases}$$

and the predicted distribution

$$Q(y|x_i)$$

Supposing that all data are independent from each other, the **likelihood of** Q **given** P is

$$\prod_{i=1}^{N} Q(y_i|x_i)$$

The likelihood of data (2)

The quantity $Q(y_i|x_i)$ can be unfolded w.r.t. the possible discrete values of $y \in \{1, \ldots m\}$:

$$Q(y_i|x_i) = \prod_{k=1}^{m} Q(y = k|x_i)^{P(y=k|x_i)}$$

Observe that the exponent is always 0 unless when $k = y_i$ (the right label of x_i), in which case the exponent is 1.

So the likelihood of Q given P can be expressed as

$$\prod_{i=1}^{N} \prod_{k=1}^{m} Q(y = k|x_i)^{P(y=k|x_i)}$$

From likelihood to negative loglikelihood

Our objective is to maximize the likelihood of the predicted distribution.

Equivalently, we can minimize the negative logarithm of the likelihood, known as negative loglikelihood:

$$-log(\prod_{i=1}^{N} \prod_{k=1}^{m} Q(y = k|x_i)^{P(y=k|x_i)})$$

and using well known properties of logarithms, this can be rewritten as

$$\sum_{i=1}^{N} - \sum_{k=1}^{m} P(y = k|x_i) \log Q(y = k|x_i)$$

negative loglikelihood and crossentropy

Given two probability distributions P and Q, the quantity

$$\mathcal{H}(P,Q) = -\sum_k P(k)\log Q(k)$$

is the **crossentropy** between P and Q.

It is a measure of the information loss due to approximating P with Q.

The croessntopy is minimal when $Q = P$, and in that case it coincides with the entropy of P:

$$\mathcal{H}(P) = -\sum_k P(k)\log P(k)$$

The negative loglikelihood is hence just the crossentropy over all data in the dataset.

Kullback-Leibler divergence

Another interesting notion of similarity between probability distributions is the so called Kullback-Leibler divergence $DKL(P\|Q)$. (it's not symmetric)

Formally,

$$
\begin{aligned}
DKL(P\|Q) &= \sum_i P(i) log \frac{P(i)}{Q(i)} \\
&= \sum_i P(i)(\log P(i) - \log Q(i)) \\
&= \sum_i P(i) \log P(i) - \sum_i P(i) \log Q(i)) \\
&= -\underbrace{\mathcal{H}(P)}_{entropy} + \underbrace{\mathcal{H}(P,Q)}_{crossentropy}
\end{aligned}
$$

Equivalently

$$\mathcal{H}(P,Q) = \mathcal{H}(P) + DKL(P\|Q)$$

Minimizing the cross entropy

Let P be the distribution of training data, and Q the distribution induced by the model.

We can take as our learning objective the minimization of the Kullback-Leibler divergence $DKL(P\|Q)$.

Since, given the training data, their entropy $\mathcal{H}(P)$ is constant, minimizing $DKL(P\|Q)$ is equivalent to minimizing the cross-entropy $\mathcal{H}(P, Q)$ between P and Q.

+

Summing up

For **binary classification** use:

- sigmoid as activation function
- binary crossentropy (aka loglikelihood) as loss function

For **multinomial classification** use:

- softmax as activation function
- categorical crossentropy as loss function

Appendix: Entropy recap

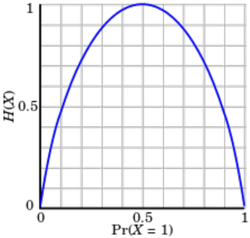
JUST READ

The entropy $H(X)$ of a random variable X is

$$H(X) = -\sum_{i=1}^{n} P(X=i) log_2 P(X=i)$$

where n is the number of possible values of X.

Entropy measures the degree of impurity of the information. It is maximal when X is uniformly distributed over all values, and minimal (0) when it is concentrated on a single value.



Information Theory (Shannon)

Entropy can be understood as the amount of information produced by a stochastic source of data.

Information is associated with the *probability* of each data (the "surprise" carried by the event):

- an event with probability 1 carries no information: $I(1) = 0$
- given two independent events with probabilities p_1 and p_2 their joint probability is $p_1 p_2$ but the information acquired is the sum of the informations of the two independent events, so

$$I(p_1 p_2) = I(p_1) + I(p_2)$$

It is hence natural to define

$$I(p) = -log(p)$$

Code Theory (Shannon)

Entropy also measures the average number of bits required to transmit outcomes produced by stochastic process X.

Suppose to have n events with the same probability. How many bits do you need to encode each possible outcome?

$$log(n)$$

In this case,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(X=i) log_2 P(X=i) \\ &= -\sum_{i=1}^{n} 1/n \, log_2(1/n) \\ &= log(n) \end{aligned}$$

If events are not equiprobable we can do better!!!

Code Theory (Shannon)

Entropy also measures the average number of bits required to transmit outcomes produced by stochastic process X.

Suppose to have n events with the same probability. How many bits do you need to encode each possible outcome?

$$log(n)$$

In this case,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(X=i) log_2 P(X=i) \\ &= -\sum_{i=1}^{n} 1/n \, log_2(1/n) \\ &= log(n) \end{aligned}$$

If events are not equiprobable we can do better!!!

Code Theory (Shannon)

Entropy also measures the average number of bits required to transmit outcomes produced by stochastic process X.

Suppose to have n events with the same probability. How many bits do you need to encode each possible outcome?

$$log(n)$$

In this case,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(X=i) log_2 P(X=i) \\ &= -\sum_{i=1}^{n} 1/n\, log_2(1/n) \\ &= log(n) \end{aligned}$$

If events are not equiprobable we can do better!!!

Code Theory (Shannon)

Entropy also measures the average number of bits required to transmit outcomes produced by stochastic process X.

Suppose to have n events with the same probability. How many bits do you need to encode each possible outcome?

$$log(n)$$

In this case,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(X=i) log_2 P(X=i) \\ &= -\sum_{i=1}^{n} 1/n\, log_2(1/n) \\ &= log(n) \end{aligned}$$

If events are not equiprobable we can do better!!!

Related notions:

Entropy of X

$$H(X) = -\sum_{i=1}^{n} P(X=i) log_2 P(X=i)$$

Conditional Entropy of X given a specific $Y = v$

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v) log_2 P(X=i|Y=v)$$

Conditional Entropy of X given Y
(weighted average over all m possible values of Y)

$$H(X|Y) = \sum_{v=1}^{m} P(Y=v) H(X|Y=v)$$

Information Gain between X and Y:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$